Extending and Embedding the Python Interpreter

Guido van Rossum
Dept. AA, CWI, P.O. Box 94079
1090 GB Amsterdam, The Netherlands
E-mail: guido@cwi.nl

10 April 1995
Release 1.2

Abstract

Python is an interpreted, object-oriented programming language. This document describes how to write modules in C or C++ to extend the Python interpreter with new modules. Those modules can define new functions but also new object types and their methods. The document also describes how to embed the Python interpreter in another application, for use as an extension language. Finally, it shows how to compile and link extension modules so that they can be loaded dynamically (at run time) into the interpreter, if the underlying operating system supports this feature.

This document assumes basic knowledge about Python. For an informal introduction to the language, see the Python Tutorial. The Python Reference Manual gives a more formal definition of the language. The Python Library Reference documents the existing object types, functions and modules (both built-in and written in Python) that give the language its wide application range.

Contents

Chapter 1

Extending Python with C or C++ code

1.1 Introduction

It is quite easy to add new built-in modules to Python, if you know how to program in C. Such *extension modules* can do two things that can't be done directly in Python: they can implement new built-in object types, and they can call C library functions and system calls.

To support extensions, the Python API (Application Programmers Interface) defines a set of functions, macros and variables that provide access to most aspects of the Python run-time system. The Python API is incorporated in a C source file by including the header `"Python.h"`.

The compilation of an extension module depends on its intended use as well as on your system setup; details are given in a later section.

1.2 A Simple Example

Let's create an extension module called '`spam`' (the favorite food of Monty Python fans...) and let's say we want to create a Python interface to the C library function `system()`.[1] This function takes a null-terminated character string as argument and returns an integer. We want this function to be callable from Python as follows:

```
>>> import spam
>>> status = spam.system("ls -l")
```

Begin by creating a file '`spammodule.c`'. (In general, if a module is called '`spam`', the C file containing its implementation is called '`spammodule.c`'; if the module name is very long, like '`spammify`', the module name can be just '`spammify.c`'.)

The first line of our file can be:

```
#include "Python.h"
```

[1] An interface for this function already exists in the standard module `os` — it was chosen as a simple and straightfoward example.

which pulls in the Python API (you can add a comment describing the purpose of the module and a copyright notice if you like).

All user-visible symbols defined by `"Python.h"` have a prefix of '`Py`' or '`PY`', except those defined in standard header files. For convenience, and since they are used extensively by the Python interpreter, `"Python.h"` includes a few standard header files: `<stdio.h>`, `<string.h>`, `<errno.h>`, and `<stdlib.h>`. If the latter header file does not exist on your system, it declares the functions `malloc()`, `free()` and `realloc()` directly.

The next thing we add to our module file is the C function that will be called when the Python expression '`spam.system(`*string*`)`' is evaluated (we'll see shortly how it ends up being called):

```
static PyObject *
spam_system(self, args)
    PyObject *self;
    PyObject *args;
{
    char *command;
    int sts;
    if (!PyArg_ParseTuple(args, "s", &command))
        return NULL;
    sts = system(command);
    return Py_BuildValue("i", sts);
}
```

There is a straightforward translation from the argument list in Python (e.g. the single expression `"ls -l"`) to the arguments passed to the C function. The C function always has two arguments, conventionally named *self* and *args*.

The *self* argument is only used when the C function implements a builtin method. This will be discussed later. In the example, *self* will always be a `NULL` pointer, since we are defining a function, not a method. (This is done so that the interpreter doesn't have to understand two different types of C functions.)

The *args* argument will be a pointer to a Python tuple object containing the arguments. Each item of the tuple corresponds to an argument in the call's argument list. The arguments are Python objects – in order to do anything with them in our C function we have to convert them to C values. The function `PyArg_ParseTuple()` in the Python API checks the argument types and converts them to C values. It uses a template string to determine the required types of the arguments as well as the types of the C variables into which to store the converted values. More about this later.

`PyArg_ParseTuple()` returns true (nonzero) if all arguments have the right type and its components have been stored in the variables whose addresses are passed. It returns false (zero) if an invalid argument list was passed. In the latter case it also raises an appropriate exception by so the calling function can return `NULL` immediately (as we saw in the example).

1.3 Intermezzo: Errors and Exceptions

An important convention throughout the Python interpreter is the following: when a function fails, it should set an exception condition and return an error value (usually a `NULL` pointer). Exceptions are stored in a static global variable inside the interpreter; if this variable is `NULL` no exception has occurred. A second global variable stores the "associated value" of the exception (the second argument to `raise`). A third variable contains the stack traceback in case the error originated in Python code. These three variables are the C equivalents of the Python variables `sys.exc_type`, `sys.exc_value` and `sys.exc_traceback` (see the section on module `sys` in the Library Reference Manual). It is important to know about them to understand how errors are passed around.

The Python API defines a number of functions to set various types of exceptions.

The most common one is `PyErr_SetString()`. Its arguments are an exception object and a C string. The exception object is usually a predefined object like `PyExc_ZeroDivisionError`. The C string indicates the cause of the error and is converted to a Python string object and stored as the "associated value" of the exception.

Another useful function is `PyErr_SetFromErrno()`, which only takes an exception argument and constructs the associated value by inspection of the (UNIX) global variable `errno`. The most general function is `PyErr_SetObject()`, which takes two object arguments, the exception and its associated value. You don't need to `Py_INCREF()` the objects passed to any of these functions.

You can test non-destructively whether an exception has been set with `PyErr_Occurred()`. This returns the current exception object, or `NULL` if no exception has occurred. You normally don't need to call `PyErr_Occurred()` to see whether an error occurred in a function call, since you should be able to tell from the return value.

When a function *f* that calls another function varg detects that the latter fails, *f* should itself return an error value (e.g. `NULL` or `-1`). It should *not* call one of the `PyErr_*()` functions — one has already been called by *g*. *f*'s caller is then supposed to also return an error indication to *its* caller, again *without* calling `PyErr_*()`, and so on — the most detailed cause of the error was already reported by the function that first detected it. Once the error reaches the Python interpreter's main loop, this aborts the currently executing Python code and tries to find an exception handler specified by the Python programmer.

(There are situations where a module can actually give a more detailed error message by calling another `PyErr_*()` function, and in such cases it is fine to do so. As a general rule, however, this is not necessary, and can cause information about the cause of the error to be lost: most operations can fail for a variety of reasons.)

To ignore an exception set by a function call that failed, the exception condition must be cleared explicitly by calling `PyErr_Clear()`. The only time C code should call `PyErr_Clear()` is if it doesn't want to pass the error on to the interpreter but wants to handle it completely by itself (e.g. by trying something else or pretending nothing happened).

Note that a failing `malloc()` call must be turned into an exception — the direct caller of `malloc()` (or `realloc()`) must call `PyErr_NoMemory()` and return a failure indicator itself. All the object-creating functions (`PyInt_FromLong()` etc.) already do this, so only if you call `malloc()` directly this note is of importance.

Also note that, with the important exception of `PyArg_ParseTuple()` and friends, functions that return an integer status usually return a positive value or zero for success and `-1` for failure, like UNIX system calls.

Finally, be careful to clean up garbage (by making `Py_XDECREF()` or `Py_DECREF()` calls for objects you have already created) when you return an error indicator!

The choice of which exception to raise is entirely yours. There are predeclared C objects corresponding to all built-in Python exceptions, e.g. `PyExc_ZeroDevisionError` which you can use directly. Of course, you should choose exceptions wisely — don't use `PyExc_TypeError` to mean that a file couldn't be opened (that should probably be `PyExc_IOError`). If something's wrong with the argument list, the `PyArg_ParseTuple()` function usually raises `PyExc_TypeError`. If you have an argument whose value which must be in a particular range or must satisfy other conditions, `PyExc_ValueError` is appropriate.

You can also define a new exception that is unique to your module. For this, you usually declare a static object variable at the beginning of your file, e.g.

```
static PyObject *SpamError;
```

and initialize it in your module's initialization function (`initspam()`) with a string object, e.g. (leaving out the error checking for now):

```
void
initspam()
{
    PyObject *m, *d;
    m = Py_InitModule("spam", SpamMethods);
    d = PyModule_GetDict(m);
    SpamError = PyString_FromString("spam.error");
    PyDict_SetItemString(d, "error", SpamError);
}
```

Note that the Python name for the exception object is `spam.error`. It is conventional for module and exception names to be spelled in lower case. It is also conventional that the *value* of the exception object is the same as its name, e.g. the string `"spam.error"`.

1.4 Back to the Example

Going back to our example function, you should now be able to understand this statement:

```
    if (!PyArg_ParseTuple(args, "s", &command))
        return NULL;
```

It returns `NULL` (the error indicator for functions returning object pointers) if an error is detected in the argument list, relying on the exception set by `PyArg_ParseTuple()`. Otherwise the string value of the argument has been copied to the local variable `command`. This is a pointer assignment and you are not supposed to modify the string to which it points (so in Standard C, the variable `command` should properly be declared as '`const char *command`').

The next statement is a call to the UNIX function `system()`, passing it the string we just got from `PyArg_ParseTuple()`:

```
    sts = system(command);
```

Our `spam.system()` function must return the value of `sys` as a Python object. This is done using the function `Py_BuildValue()`, which is something like the inverse of `PyArg_ParseTuple()`: it takes a format string and an arbitrary number of C values, and returns a new Python object. More info on `Py_BuildValue()` is given later.

```
    return Py_BuildValue("i", sts);
```

In this case, it will return an integer object. (Yes, even integers are objects on the heap in Python!)

If you have a C function that returns no useful argument (a function returning `void`), the corresponding Python function must return `None`. You need this idiom to do so:

```
    Py_INCREF(Py_None);
    return Py_None;
```

`Py_None` is the C name for the special Python object `None`. It is a genuine Python object (not a `NULL` pointer, which means "error" in most contexts, as we have seen).

1.5 The Module's Method Table and Initialization Function

I promised to show how `spam_system()` is called from Python programs. First, we need to list its name and address in a "method table":

```
static PyMethodDef SpamMethods[] = {
    ...
    {"system",  spam_system, 1},
    ...
    {NULL,      NULL}        /* Sentinel */
};
```

Note the third entry ('`1`'). This is a flag telling the interpreter the calling convention to be used for the C function. It should normally always be '`1`'; a value of '`0`' means that an obsolete variant of `PyArg_ParseTuple()` is used.

The method table must be passed to the interpreter in the module's initialization function (which should be the only non-`static` item defined in the module file):

```
void
initspam()
{
    (void) Py_InitModule("spam", SpamMethods);
}
```

When the Python program imports module `spam` for the first time, `initspam()` is called. It calls `Py_InitModule()`, which creates a "module object" (which is inserted in the dictionary `sys.modules` under the key `"spam"`), and inserts built-in function objects into the newly created module based upon the table (an array of `PyMethodDef` structures) that was passed as its second argument. `Py_InitModule()` returns a pointer to the module object that it creates (which is unused here). It aborts with a fatal error if the module could not be initialized satisfactorily, so the caller doesn't need to check for errors.

1.6 Compilation and Linkage

There are two more things to do before you can use your new extension: compiling and linking it with the Python system. If you use dynamic loading, the details depend on the style of dynamic loading your system uses; see the chapter on Dynamic Loading for more info about this.

If you can't use dynamic loading, or if you want to make your module a permanent part of the Python interpreter, you will have to change the configuration setup and rebuild the interpreter. Luckily, this is very simple: just place your file ('`spammodule.c`' for example) in the '`Modules`' directory, add a line to the file '`Modules/Setup`' describing your file:

```
spam spammodule.o
```

and rebuild the interpreter by running `make` in the toplevel directory. You can also run `make` in the '`Modules`' subdirectory, but then you must first rebuilt the '`Makefile`' there by running `make Makefile`. (This is necessary each time you change the '`Setup`' file.)

If your module requires additional libraries to link with, these can be listed on the line in the '`Setup`' file as well, for instance:

```
spam spammodule.o -lX11
```

1.7 Calling Python Functions From C

So far we have concentrated on making C functions callable from Python. The reverse is also useful: calling Python functions from C. This is especially the case for libraries that support so-called "callback" functions. If a C interface makes use of callbacks, the equivalent Python often needs to provide a callback mechanism to the Python programmer; the implementation will require calling the Python callback functions from a C callback. Other uses are also imaginable.

Fortunately, the Python interpreter is easily called recursively, and there is a standard interface to call a Python function. (I won't dwell on how to call the Python parser with a particular string as input — if you're interested, have a look at the implementation of the '-c' command line option in 'Python/pythonmain.c'.)

Calling a Python function is easy. First, the Python program must somehow pass you the Python function object. You should provide a function (or some other interface) to do this. When this function is called, save a pointer to the Python function object (be careful to Py_INCREF() it!) in a global variable — or whereever you see fit. For example, the following function might be part of a module definition:

```
static PyObject *my_callback = NULL;

static PyObject *
my_set_callback(dummy, arg)
    PyObject *dummy, *arg;
{
    Py_XDECREF(my_callback); /* Dispose of previous callback */
    Py_XINCREF(arg); /* Add a reference to new callback */
    my_callback = arg; /* Remember new callback */
    /* Boilerplate to return "None" */
    Py_INCREF(Py_None);
    return Py_None;
}
```

The macros Py_XINCREF() and Py_XDECREF() increment/decrement the reference count of an object and are safe in the presence of NULL pointers. More info on them in the section on Reference Counts below.

Later, when it is time to call the function, you call the C function PyEval_CallObject(). This function has two arguments, both pointers to arbitrary Python objects: the Python function, and the argument list. The argument list must always be a tuple object, whose length is the number of arguments. To call the Python function with no arguments, pass an empty tuple; to call it with one argument, pass a singleton tuple. Py_BuildValue() returns a tuple when its format string consists of zero or more format codes between parentheses. For example:

```
int arg;
PyObject *arglist;
PyObject *result;
...
arg = 123;
...
/* Time to call the callback */
arglist = Py_BuildValue("(i)", arg);
result = PyEval_CallObject(my_callback, arglist);
Py_DECREF(arglist);
```

PyEval_CallObject() returns a Python object pointer: this is the return value of the Python function. PyEval_CallObject() is "reference-count-neutral" with respect to its arguments. In the example a new tuple was created to serve as the argument list, which is Py_DECREF()-ed immediately after the call.

The return value of PyEval_CallObject() is “new”: either it is a brand new object, or it is an existing object whose reference count has been incremented. So, unless you want to save it in a global variable, you should somehow Py_DECREF() the result, even (especially!) if you are not interested in its value.

Before you do this, however, it is important to check that the return value isn't NULL. If it is, the Python function terminated by raising an exception. If the C code that called PyEval_CallObject() is called from Python, it should now return an error indication to its Python caller, so the interpreter can print a stack trace, or the calling Python code can handle the exception. If this is not possible or desirable, the exception should be cleared by calling PyErr_Clear(). For example:

```
if (result == NULL)
    return NULL; /* Pass error back */
...use result...
Py_DECREF(result);
```

Depending on the desired interface to the Python callback function, you may also have to provide an argument list to PyEval_CallObject(). In some cases the argument list is also provided by the Python program, through the same interface that specified the callback function. It can then be saved and used in the same manner as the function object. In other cases, you may have to construct a new tuple to pass as the argument list. The simplest way to do this is to call Py_BuildValue(). For example, if you want to pass an integral event code, you might use the following code:

```
PyObject *arglist;
...
arglist = Py_BuildValue("(l)", eventcode);
result = PyEval_CallObject(my_callback, arglist);
Py_DECREF(arglist);
if (result == NULL)
    return NULL; /* Pass error back */
/* Here maybe use the result */
Py_DECREF(result);
```

Note the placement of Py_DECREF(argument) immediately after the call, before the error check! Also note that strictly spoken this code is not complete: Py_BuildValue() may run out of memory, and this should be checked.

1.8 Format Strings for PyArg_ParseTuple()

The PyArg_ParseTuple() function is declared as follows:

```
int PyArg_ParseTuple(PyObject *arg, char *format, ...);
```

The *arg* argument must be a tuple object containing an argument list passed from Python to a C function. The *format* argument must be a format string, whose syntax is explained below. The remaining arguments must be addresses of variables whose type is determined by the format string. For the conversion to succeed, the *arg* object must match the format and the format must be exhausted.

Note that while `PyArg_ParseTuple()` checks that the Python arguments have the required types, it cannot check the validity of the addresses of C variables passed to the call: if you make mistakes there, your code will probably crash or at least overwrite random bits in memory. So be careful!

A format string consists of zero or more "format units". A format unit describes one Python object; it is usually a single character or a parenthesized sequence of format units. With a few exceptions, a format unit that is not a parenthesized sequence normally corresponds to a single address argument to `PyArg_ParseTuple()`. In the following description, the quoted form is the format unit; the entry in (round) parentheses is the Python object type that matches the format unit; and the entry in [square] brackets is the type of the C variable(s) whose address should be passed. (Use the '`&`' operator to pass a variable's address.)

'`s`' (string) [char *] Convert a Python string to a C pointer to a character string. You must not provide storage for the string itself; a pointer to an existing string is stored into the character pointer variable whose address you pass. The C string is null-terminated. The Python string must not contain embedded null bytes; if it does, a `TypeError` exception is raised.

'`s#`' (string) [char *, int] This variant on `'s'` stores into two C variables, the first one a pointer to a character string, the second one its length. In this case the Python string may contain embedded null bytes.

'`z`' (string or `None`) [char *] Like '`s`', but the Python object may also be `None`, in which case the C pointer is set to `NULL`.

'`z#`' (string or `None`) [char *, int] This is to `'s#'` as `'z'` is to `'s'`.

'`b`' (integer) [char] Convert a Python integer to a tiny int, stored in a C `char`.

'`h`' (integer) [short int] Convert a Python integer to a C `short int`.

'`i`' (integer) [int] Convert a Python integer to a plain C `int`.

'`l`' (integer) [long int] Convert a Python integer to a C `long int`.

'`c`' (string of length 1) [char] Convert a Python character, represented as a string of length 1, to a C `char`.

'`f`' (float) [float] Convert a Python floating point number to a C `float`.

'`d`' (float) [double] Convert a Python floating point number to a C `double`.

'`O`' (object) [PyObject *] Store a Python object (without any conversion) in a C object pointer. The C program thus receives the actual object that was passed. The object's reference count is not increased. The pointer stored is not `NULL`.

'`O!`' (object) [*typeobject*, PyObject *] Store a Python object in a C object pointer. This is similar to '`O`', but takes two C arguments: the first is the address of a Python type object, the second is the address of the C variable (of type `PyObject *`) into which the object pointer is stored. If the Python object does not have the required type, a `TypeError` exception is raised.

'`O&`' (object) [*converter*, *anything*] Convert a Python object to a C variable through a *converter* function. This takes two arguments: the first is a function, the second is the address of a C variable (of arbitrary type), converted to `void *`. The *converter* function in turn is called as follows:

status `=` *converter*`(`*object*`,` *address*`);`

where *object* is the Python object to be converted and *address* is the `void *` argument that was passed to `PyArg_ConvertTuple()`. The returned *status* should be `1` for a successful conversion and `0` if the conversion has failed. When the conversion fails, the *converter* function should raise an exception.

'`S`' (string) [PyStringObject *] Like '`O`' but raises a `TypeError` exception that the object is a string object. The C variable may also be declared as `PyObject *`.

'`(`*items*`)`' (tuple) [*matching-items*] The object must be a Python tuple whose length is the number of format units in *items*. The C arguments must correspond to the individual format units in *items*. Format units for tuples may be nested.

It is possible to pass Python long integers where integers are requested; however no proper range checking is done – the most significant bits are silently truncated when the receiving field is too small to receive the value (actually, the semantics are inherited from downcasts in C — your milage may vary).

A few other characters have a meaning in a format string. These may not occur inside nested parentheses. They are:

'`|`' Indicates that the remaining arguments in the Python argument list are optional. The C variables corresponding to optional arguments should be initialized to their default value — when an optional argument is not specified, the `PyArg_ParseTuple` does not touch the contents of the corresponding C variable(s).

'`:`' The list of format units ends here; the string after the colon is used as the function name in error messages (the "associated value" of the exceptions that `PyArg_ParseTuple` raises).

'`;`' The list of format units ends here; the string after the colon is used as the error message *instead* of the default error message. Clearly, '`:`' and '`;`' mutually exclude each other.

Some example calls:

```
int ok;
int i, j;
long k, l;
char *s;
```

```
int size;

ok = PyArg_ParseTuple(args, ""); /* No arguments */
    /* Python call: f() */

ok = PyArg_ParseTuple(args, "s", &s); /* A string */
    /* Possible Python call: f('whoops!') */

ok = PyArg_ParseTuple(args, "lls", &k, &l, &s); /* Two longs and a string */
    /* Possible Python call: f(1, 2, 'three') */

ok = PyArg_ParseTuple(args, "(ii)s#", &i, &j, &s, &size);
    /* A pair of ints and a string, whose size is also returned */
    /* Possible Python call: f(1, 2, 'three') */

{
    char *file;
    char *mode = "r";
    int bufsize = 0;
    ok = PyArg_ParseTuple(args, "s|si", &file, &mode, &bufsize);
    /* A string, and optionally another string and an integer */
    /* Possible Python calls:
       f('spam')
       f('spam', 'w')
       f('spam', 'wb', 100000) */
}

{
    int left, top, right, bottom, h, v;
    ok = PyArg_ParseTuple(args, "((ii)(ii))(ii)",
             &left, &top, &right, &bottom, &h, &v);
             /* A rectangle and a point */
             /* Possible Python call:
                f(((0, 0), (400, 300)), (10, 10)) */
}
```

1.9 The `Py_BuildValue()` Function

This function is the counterpart to `PyArg_ParseTuple()`. It is declared as follows:

```
PyObject *Py_BuildValue(char *format, ...);
```

It recognizes a set of format units similar to the ones recognized by `PyArg_ParseTuple()`, but the arguments (which are input to the function, not output) must not be pointers, just values. It returns a new Python object, suitable for returning from a C function called from Python.

One difference with `PyArg_ParseTuple()`: while the latter requires its first argument to be a tuple (since Python argument lists are always represented as tuples internally), `BuildValue()` does not always build a tuple. It builds a tuple only if its format string contains two or more format units. If the format string is empty, it returns `None`; if it contains exactly one format unit, it returns whatever object is described by that format unit. To force it to return a tuple of size 0 or one, parenthesize the format string.

In the following description, the quoted form is the format unit; the entry in (round) parentheses is the Python object type that the format unit will return; and the entry in [square] brackets is the type of the C value(s) to be passed.

The characters space, tab, colon and comma are ignored in format strings (but not within format units such as '`s#`'). This can be used to make long format strings a tad more readable.

'`s`' (string) [char *] Convert a null-terminated C string to a Python object. If the C string pointer is `NULL`, `None` is returned.

'`s#`' (string) [char *, int] Convert a C string and its length to a Python object. If the C string pointer is `NULL`, the length is ignored and `None` is returned.

'`z`' (string or `None`) [char *] Same as '`s`'.

'`z#`' (string or `None`) [char *, int] Same as '`s#`'.

'`i`' (integer) [int] Convert a plain C `int` to a Python integer object.

'`b`' (integer) [char] Same as '`i`'.

'`h`' (integer) [short int] Same as '`i`'.

'`l`' (integer) [long int] Convert a C `long int` to a Python integer object.

'`c`' (string of length 1) [char] Convert a C `int` representing a character to a Python string of length 1.

'`d`' (float) [double] Convert a C `double` to a Python floating point number.

'`f`' (float) [float] Same as '`d`'.

'`O`' (object) [PyObject *] Pass a Python object untouched (except for its reference count, which is incremented by one). If the object passed in is a `NULL` pointer, it is assumed that this was caused because the call producing the argument found an error and set an exception. Therefore, `Py_BuildValue()` will return `NULL` but won't raise an exception. If no exception has been raised yet, `PyExc_SystemError` is set.

'`S`' (object) [PyObject *] Same as '`O`'.

'`O&`' (object) [*converter*, *anything*] Convert *anything* to a Python object through a *converter* function. The function is called with *anything* (which should be compatible with `void *`) as its argument and should return a "new" Python object, or `NULL` if an error occurred.

'`(`*items*`)`' (tuple) [*matching-items*] Convert a sequence of C values to a Python tuple with the same number of items.

'`[`*items*`]`' **(list) [***matching-items***]** Convert a sequence of C values to a Python list with the same number of items.

'`{`*items*`}`' **(dictionary) [***matching-items***]** Convert a sequence of C values to a Python dictionary. Each pair of consecutive C values adds one item to the dictionary, serving as key and value, respectively.

If there is an error in the format string, the `PyExc_SystemError` exception is raised and `NULL` returned.

Examples (to the left the call, to the right the resulting Python value):

```
Py_BuildValue("")                        None
Py_BuildValue("i", 123)                  123
Py_BuildValue("iii", 123, 456, 789)      (123, 456, 789)
Py_BuildValue("s", "hello")              'hello'
Py_BuildValue("ss", "hello", "world")    ('hello', 'world')
Py_BuildValue("s#", "hello", 4)          'hell'
Py_BuildValue("()")                      ()
Py_BuildValue("(i)", 123)                (123,)
Py_BuildValue("(ii)", 123, 456)          (123, 456)
Py_BuildValue("(i,i)", 123, 456)         (123, 456)
Py_BuildValue("[i,i]", 123, 456)         [123, 456]
Py_BuildValue("{s:i,s:i}",
              "abc", 123, "def", 456)    {'abc': 123, 'def': 456}
Py_BuildValue("((ii)(ii)) (ii)",
              1, 2, 3, 4, 5, 6)          (((1, 2), (3, 4)), (5, 6))
```

1.10 Reference Counts

1.10.1 Introduction

In languages like C or C++, the programmer is responsible for dynamic allocation and deallocation of memory on the heap. In C, this is done using the functions `malloc()` and `free()`. In C++, the operators `new` and `delete` are used with essentially the same meaning; they are actually implemented using `malloc()` and `free()`, so we'll restrict the following discussion to the latter.

Every block of memory allocated with `malloc()` should eventually be returned to the pool of available memory by exactly one call to `free()`. It is important to call `free()` at the right time. If a block's address is forgotten but `free()` is not called for it, the memory it occupies cannot be reused until the program terminates. This is called a *memory leak*. On the other hand, if a program calls `free()` for a block and then continues to use the block, it creates a conflict with re-use of the block through another `malloc()` call. This is called *using freed memory* has the same bad consequences as referencing uninitialized data — core dumps, wrong results, mysterious crashes.

Common causes of memory leaks are unusual paths through the code. For instance, a function may allocate a block of memory, do some calculation, and then free the block again. Now a change in the requirements for the function may add a test to the calculation that detects an error condition and can return prematurely from the function. It's easy to forget to free the allocated memory block when taking this premature exit, especially when it is added later to the code. Such leaks, once introduced, often go undetected for a long time: the error exit is taken only in a small fraction of all calls, and most modern machines have plenty of virtual memory, so the leak only becomes apparent in a long-running process that uses the leaking function frequently. Therefore, it's important to prevent leaks from happening by having a coding convention or strategy that minimizes this kind of errors.

Since Python makes heavy use of `malloc()` and `free()`, it needs a strategy to avoid memory leaks as well as the use of freed memory. The chosen method is called *reference counting*. The principle is simple: every object contains a counter, which is incremented when a reference to the object is stored somewhere, and which is decremented when a reference to it is deleted. When the counter reaches zero, the last reference to the object has been deleted and the object is freed.

An alternative strategy is called *automatic garbage collection*. (Sometimes, reference counting is also referred to as a garbage collection strategy, hence my use of "automatic" to distinguish the two.) The big advantage of automatic garbage collection is that the user doesn't need to call `free()` explicitly. (Another claimed advantage is an improvement in speed or memory usage — this is no hard fact however.) The disadvantage is that for C, there is no truly portable automatic garbage collector, while reference counting can be implemented portably (as long as the functions `malloc()` and `free()` are available — which the C Standard guarantees). Maybe some day a sufficiently portable automatic garbage collector will be available for C. Until then, we'll have to live with reference counts.

1.10.2 Reference Counting in Python

There are two macros, `Py_INCREF(x)` and `Py_DECREF(x)`, which handle the incrementing and decrementing of the reference count. `Py_DECREF()` also frees the object when the count reaches zero. For flexibility, it doesn't call `free()` directly — rather, it makes a call through a function pointer in the object's *type object*. For this purpose (and others), every object also contains a pointer to its type object.

The big question now remains: when to use `Py_INCREF(x)` and `Py_DECREF(x)`? Let's first introduce some terms. Nobody "owns" an object; however, you can *own a reference* to an object. An object's reference count is now defined as the number of owned references to it. The owner of a reference is responsible for calling `Py_DECREF()` when the reference is no longer needed. Ownership of a reference can be transferred. There are three ways to dispose of an owned reference: pass it on, store it, or call `Py_DECREF()`. Forgetting to dispose of an owned reference creates a memory leak.

It is also possible to *borrow*[2] a reference to an object. The borrower of a reference should not call `Py_DECREF()`. The borrower must not hold on to the object longer than the owner from which it was borrowed. Using a borrowed reference after the owner has disposed of it risks using freed memory and should be avoided completely.[3]

The advantage of borrowing over owning a reference is that you don't need to take care of disposing of the reference on all possible paths through the code — in other words, with a borrowed reference

[2] The metaphor of "borrowing" a reference is not completely correct: the owner still has a copy of the reference.

[3] Checking that the reference count is at least 1 **does not work** — the reference count itself could be in freed memory and may thus be reused for another object!

you don't run the risk of leaking when a premature exit is taken. The disadvantage of borrowing over leaking is that there are some subtle situations where in seemingly correct code a borrowed reference can be used after the owner from which it was borrowed has in fact disposed of it.

A borrowed reference can be changed into an owned reference by calling `Py_INCREF()`. This does not affect the status of the owner from which the reference was borrowed — it creates a new owned reference, and gives full owner responsibilities (i.e., the new owner must dispose of the reference properly, as well as the previous owner).

1.10.3 Ownership Rules

Whenever an object reference is passed into or out of a function, it is part of the function's interface specification whether ownership is transferred with the reference or not.

Most functions that return a reference to an object pass on ownership with the reference. In particular, all functions whose function it is to create a new object, e.g. `PyInt_FromLong()` and `Py_BuildValue()`, pass ownership to the receiver. Even if in fact, in some cases, you don't receive a reference to a brand new object, you still receive ownership of the reference. For instance, `PyInt_FromLong()` maintains a cache of popular values and can return a reference to a cached item.

Many functions that extract objects from other objects also transfer ownership with the reference, for instance `PyObject_GetAttrString()`. The picture is less clear, here, however, since a few common routines are exceptions: `PyTuple_GetItem()`, `PyList_GetItem()` and `PyDict_GetItem()` (and `PyDict_GetItemString()`) all return references that you borrow from the tuple, list or dictionary.

The function `PyImport_AddModule()` also returns a borrowed reference, even though it may actually create the object it returns: this is possible because an owned reference to the object is stored in `sys.modules`.

When you pass an object reference into another function, in general, the function borrows the reference from you — if it needs to store it, it will use `Py_INCREF()` to become an independent owner. There are exactly two important exceptions to this rule: `PyTuple_SetItem()` and `PyList_SetItem()`. These functions take over ownership of the item passed to them — even if they fail! (Note that `PyDict_SetItem()` and friends don't take over ownership — they are "normal".)

When a C function is called from Python, it borrows references to its arguments from the caller. The caller owns a reference to the object, so the borrowed reference's lifetime is guaranteed until the function returns. Only when such a borrowed reference must be stored or passed on, it must be turned into an owned reference by calling `Py_INCREF()`.

The object reference returned from a C function that is called from Python must be an owned reference — ownership is tranferred from the function to its caller.

1.10.4 Thin Ice

There are a few situations where seemingly harmless use of a borrowed reference can lead to problems. These all have to do with implicit invocations of the interpreter, which can cause the owner of a reference to dispose of it.

The first and most important case to know about is using `Py_DECREF()` on an unrelated object while borrowing a reference to a list item. For instance:

```
bug(PyObject *list) {
    PyObject *item = PyList_GetItem(list, 0);
    PyList_SetItem(list, 1, PyInt_FromLong(0L));
    PyObject_Print(item, stdout, 0); /* BUG! */
}
```

This function first borrows a reference to `list[0]`, then replaces `list[1]` with the value `0`, and finally prints the borrowed reference. Looks harmless, right? But it's not!

Let's follow the control flow into `PyList_SetItem()`. The list owns references to all its items, so when item 1 is replaced, it has to dispose of the original item 1. Now let's suppose the original item 1 was an instance of a user-defined class, and let's further suppose that the class defined a `__del__()` method. If this class instance has a reference count of 1, disposing of it will call its `__del__()` method.

Since it is written in Python, the `__del__()` method can execute arbitrary Python code. Could it perhaps do something to invalidate the reference to `item` in `bug()`? You bet! Assuming that the list passed into `bug()` is accessible to the `__del__()` method, it could execute a statement to the effect of `del list[0]`, and assuming this was the last reference to that object, it would free the memory associated with it, thereby invalidating `item`.

The solution, once you know the source of the problem, is easy: temporarily increment the reference count. The correct version of the function reads:

```
no_bug(PyObject *list) {
    PyObject *item = PyList_GetItem(list, 0);
    Py_INCREF(item);
    PyList_SetItem(list, 1, PyInt_FromLong(0L));
    PyObject_Print(item, stdout, 0);
    Py_DECREF(item);
}
```

This is a true story. An older version of Python contained variants of this bug and someone spent a considerable amount of time in a C debugger to figure out why his `__del__()` methods would fail...

The second case of problems with a borrowed reference is a variant involving threads. Normally, multiple threads in the Python interpreter can't get in each other's way, because there is a global lock protecting Python's entire object space. However, it is possible to temporarily release this lock using the macro `Py_BEGIN_ALLOW_THREADS`, and to re-acquire it using `Py_END_ALLOW_THREADS`. This is common around blocking I/O calls, to let other threads use the CPU while waiting for the I/O to complete. Obviously, the following function has the same problem as the previous one:

```
bug(PyObject *list) {
    PyObject *item = PyList_GetItem(list, 0);
```

```
    Py_BEGIN_ALLOW_THREADS
    ...some blocking I/O call...
    Py_END_ALLOW_THREADS
    PyObject_Print(item, stdout, 0); /* BUG! */
}
```

1.10.5 NULL Pointers

In general, functions that take object references as arguments don't expect you to pass them `NULL` pointers, and will dump core (or cause later core dumps) if you do so. Functions that return object references generally return `NULL` only to indicate that an exception occurred. The reason for not testing for `NULL` arguments is that functions often pass the objects they receive on to other function — if each function were to test for `NULL`, there would be a lot of redundant tests and the code would run slower.

It is better to test for `NULL` only at the "source", i.e. when a pointer that may be `NULL` is received, e.g. from `malloc()` or from a function that may raise an exception.

The macros `Py_INCREF()` and `Py_DECREF()` don't check for `NULL` pointers — however, their variants `Py_XINCREF()` and `Py_XDECREF()` do.

The macros for checking for a particular object type (`Py`*type*`_Check()`) don't check for `NULL` pointers — again, there is much code that calls several of these in a row to test an object against various different expected types, and this would generate redundant tests. There are no variants with `NULL` checking.

The C function calling mechanism guarantees that the argument list passed to C functions (`args` in the examples) is never `NULL` — in fact it guarantees that it is always a tuple.[4]

It is a severe error to ever let a `NULL` pointer "escape" to the Python user.

1.11 Writing Extensions in C++

It is possible to write extension modules in C++. Some restrictions apply: since the main program (the Python interpreter) is compiled and linked by the C compiler, global or static objects with constructors cannot be used. All functions that will be called directly or indirectly (i.e. via function pointers) by the Python interpreter will have to be declared using `extern "C"`; this applies to all "methods" as well as to the module's initialization function. It is unnecessary to enclose the Python header files in `extern "C" {...}` — they use this form already if the symbol '`__cplusplus`' is defined (all recent C++ compilers define this symbol).

[4]These guarantees don't hold when you use the "old" style calling convention — this is still found in much existing code.

Chapter 2

Embedding Python in another application

Embedding Python is similar to extending it, but not quite. The difference is that when you extend Python, the main program of the application is still the Python interpreter, while if you embed Python, the main program may have nothing to do with Python — instead, some parts of the application occasionally call the Python interpreter to run some Python code.

So if you are embedding Python, you are providing your own main program. One of the things this main program has to do is initialize the Python interpreter. At the very least, you have to call the function `Py_Initialize()`. There are optional calls to pass command line arguments to Python. Then later you can call the interpreter from any part of the application.

There are several different ways to call the interpreter: you can pass a string containing Python statements to `PyRun_SimpleString()`, or you can pass a stdio file pointer and a file name (for identification in error messages only) to `PyRun_SimpleFile()`. You can also call the lower-level operations described in the previous chapters to construct and use Python objects.

A simple demo of embedding Python can be found in the directory '`Demo/embed`'.

2.1 Embedding Python in C++

It is also possible to embed Python in a C++ program; precisely how this is done will depend on the details of the C++ system used; in general you will need to write the main program in C++, and use the C++ compiler to compile and link your program. There is no need to recompile Python itself using C++.

Chapter 3

Dynamic Loading

On most modern systems it is possible to configure Python to support dynamic loading of extension modules implemented in C. When shared libraries are used dynamic loading is configured automatically; otherwise you have to select it as a build option (see below). Once configured, dynamic loading is trivial to use: when a Python program executes `import spam`, the search for modules tries to find a file '`spammodule.o`' ('`spammodule.so`' when using shared libraries) in the module search path, and if one is found, it is loaded into the executing binary and executed. Once loaded, the module acts just like a built-in extension module.

The advantages of dynamic loading are twofold: the "core" Python binary gets smaller, and users can extend Python with their own modules implemented in C without having to build and maintain their own copy of the Python interpreter. There are also disadvantages: dynamic loading isn't available on all systems (this just means that on some systems you have to use static loading), and dynamically loading a module that was compiled for a different version of Python (e.g. with a different representation of objects) may dump core.

3.1 Configuring and Building the Interpreter for Dynamic Loading

There are three styles of dynamic loading: one using shared libraries, one using SGI IRIX 4 dynamic loading, and one using GNU dynamic loading.

3.1.1 Shared Libraries

The following systems support dynamic loading using shared libraries: SunOS 4; Solaris 2; SGI IRIX 5 (but not SGI IRIX 4!); and probably all systems derived from SVR4, or at least those SVR4 derivatives that support shared libraries (are there any that don't?).

You don't need to do anything to configure dynamic loading on these systems — the '`configure`' detects the presence of the '`<dlfcn.h>`' header file and automatically configures dynamic loading.

3.1.2 SGI IRIX 4 Dynamic Loading

Only SGI IRIX 4 supports dynamic loading of modules using SGI dynamic loading. (SGI IRIX 5 might also support it but it is inferior to using shared libraries so there is no reason to; a small test didn't work right away so I gave up trying to support it.)

Before you build Python, you first need to fetch and build the `dl` package written by Jack Jansen. This is available by anonymous ftp from host '`ftp.cwi.nl`', directory '`pub/dynload`', file '`dl-1.6.tar.Z`'. (The version number may change.) Follow the instructions in the package's '`README`' file to build it.

Once you have built `dl`, you can configure Python to use it. To this end, you run the '`configure`' script with the option `--with-dl=`*directory* where *directory* is the absolute pathname of the `dl` directory.

Now build and install Python as you normally would (see the '`README`' file in the toplevel Python directory.)

3.1.3 GNU Dynamic Loading

GNU dynamic loading supports (according to its '`README`' file) the following hardware and software combinations: VAX (Ultrix), Sun 3 (SunOS 3.4 and 4.0), Sparc (SunOS 4.0), Sequent Symmetry (Dynix), and Atari ST. There is no reason to use it on a Sparc; I haven't seen a Sun 3 for years so I don't know if these have shared libraries or not.

You need to fetch and build two packages. One is GNU DLD 3.2.3, available by anonymous ftp from host '`ftp.cwi.nl`', directory '`pub/dynload`', file '`dld-3.2.3.tar.Z`'. (As far as I know, no further development on GNU DLD is being done.) The other is an emulation of Jack Jansen's `dl` package that I wrote on top of GNU DLD 3.2.3. This is available from the same host and directory, file dl-dld-1.1.tar.Z. (The version number may change — but I doubt it will.) Follow the instructions in each package's '`README`' file to configure build them.

Now configure Python. Run the '`configure`' script with the option `--with-dl-dld=`*dl-directory*`,`*dld-directory* where *dl-directory* is the absolute pathname of the directory where you have built the '`dl-dld`' package, and *dld-directory* is that of the GNU DLD package. The Python interpreter you build hereafter will support GNU dynamic loading.

3.2 Building a Dynamically Loadable Module

Since there are three styles of dynamic loading, there are also three groups of instructions for building a dynamically loadable module. Instructions common for all three styles are given first. Assuming your module is called `spam`, the source filename must be '`spammodule.c`', so the object name is '`spammodule.o`'. The module must be written as a normal Python extension module (as described earlier).

Note that in all cases you will have to create your own Makefile that compiles your module file(s). This Makefile will have to pass two '`-I`' arguments to the C compiler which will make it find the Python header files. If the Make variable *PYTHONTOP* points to the toplevel Python directory, your *CFLAGS*

Make variable should contain the options '`-I$(PYTHONTOP) -I$(PYTHONTOP)/Include`'. (Most header files are in the '`Include`' subdirectory, but the '`config.h`' header lives in the toplevel directory.) You must also add '`-DHAVE_CONFIG_H`' to the definition of *CFLAGS* to direct the Python headers to include '`config.h`'.

3.2.1 Shared Libraries

You must link the '`.o`' file to produce a shared library. This is done using a special invocation of the UNIX loader/linker, *ld*(1). Unfortunately the invocation differs slightly per system.

On SunOS 4, use

```
ld spammodule.o -o spammodule.so
```

On Solaris 2, use

```
ld -G spammodule.o -o spammodule.so
```

On SGI IRIX 5, use

```
ld -shared spammodule.o -o spammodule.so
```

On other systems, consult the manual page for `ld`(1) to find what flags, if any, must be used.

If your extension module uses system libraries that haven't already been linked with Python (e.g. a windowing system), these must be passed to the `ld` command as '`-l`' options after the '`.o`' file.

The resulting file '`spammodule.so`' must be copied into a directory along the Python module search path.

3.2.2 SGI IRIX 4 Dynamic Loading

IMPORTANT: You must compile your extension module with the additional C flag '`-G0`' (or '`-G 0`'). This instruct the assembler to generate position-independent code.

You don't need to link the resulting '`spammodule.o`' file; just copy it into a directory along the Python module search path.

The first time your extension is loaded, it takes some extra time and a few messages may be printed. This creates a file '`spammodule.ld`' which is an image that can be loaded quickly into the Python interpreter process. When a new Python interpreter is installed, the `dl` package detects this and rebuilds '`spammodule.ld`'. The file '`spammodule.ld`' is placed in the directory where '`spammodule.o`' was found, unless this directory is unwritable; in that case it is placed in a temporary directory.[1]

If your extension modules uses additional system libraries, you must create a file '`spammodule.libs`' in the same directory as the '`spammodule.o`'. This file should contain one or more lines with whitespace-separated options that will be passed to the linker — normally only '`-l`' options or absolute pathnames of libraries ('`.a`' files) should be used.

[1]Check the manual page of the `dl` package for details.

3.2.3 GNU Dynamic Loading

Just copy 'spammodule.o' into a directory along the Python module search path.

If your extension modules uses additional system libraries, you must create a file 'spammodule.libs' in the same directory as the 'spammodule.o'. This file should contain one or more lines with whitespace-separated absolute pathnames of libraries ('.a' files). No '-l' options can be used.